Large Language Models Exhibit Human-Like Bayesian Hypocrisy

Nykko Vitali & Mahzarin R. Banaji
Department of Psychology, Harvard University

Introduction

Previous work by Cao et al. (2019) demonstrated humans exhibit accurate Bayesian reasoning yet condemn others who make identical statistical judgments, a phenomenon termed **Bayesian Hypocrisy**

As LLMs are deployed in sensitive decision-making domains, we tested whether they replicate this human inconsistency between statistical reasoning and social judgment.

Research Questions

1. Do LLMs integrate priors and likelihoods to update posteriors according to Bayes' Rule?

2. Do LLMs negatively evaluate a person who makes the same Bayesian judgment they did?

Participants

Human data: 348 participants (Cao et al., 2019).

LLM data: 840 trials each from GPT-4o and Claude 3.7 Sonnet (temperature = 1.0)

Procedure

Part 1 (Pilot Scenario):

- Provide the likelihood that a flight attendant, or a pilot communicates with air traffic control (ATC).
- Provide priors for pilots and flight attendants for men and women.
- Provide a posterior given they communicated with ATC that they're the pilot (Figure 1).

Part 2 (Doctor Scenario):

- Provide a categorical assessment of who is more likely to be a doctor? Man, woman, or equally likely
- Evaluate "Person X" who stated a man is more likely to be a doctor than a woman after both performed surgery (Figure 2).

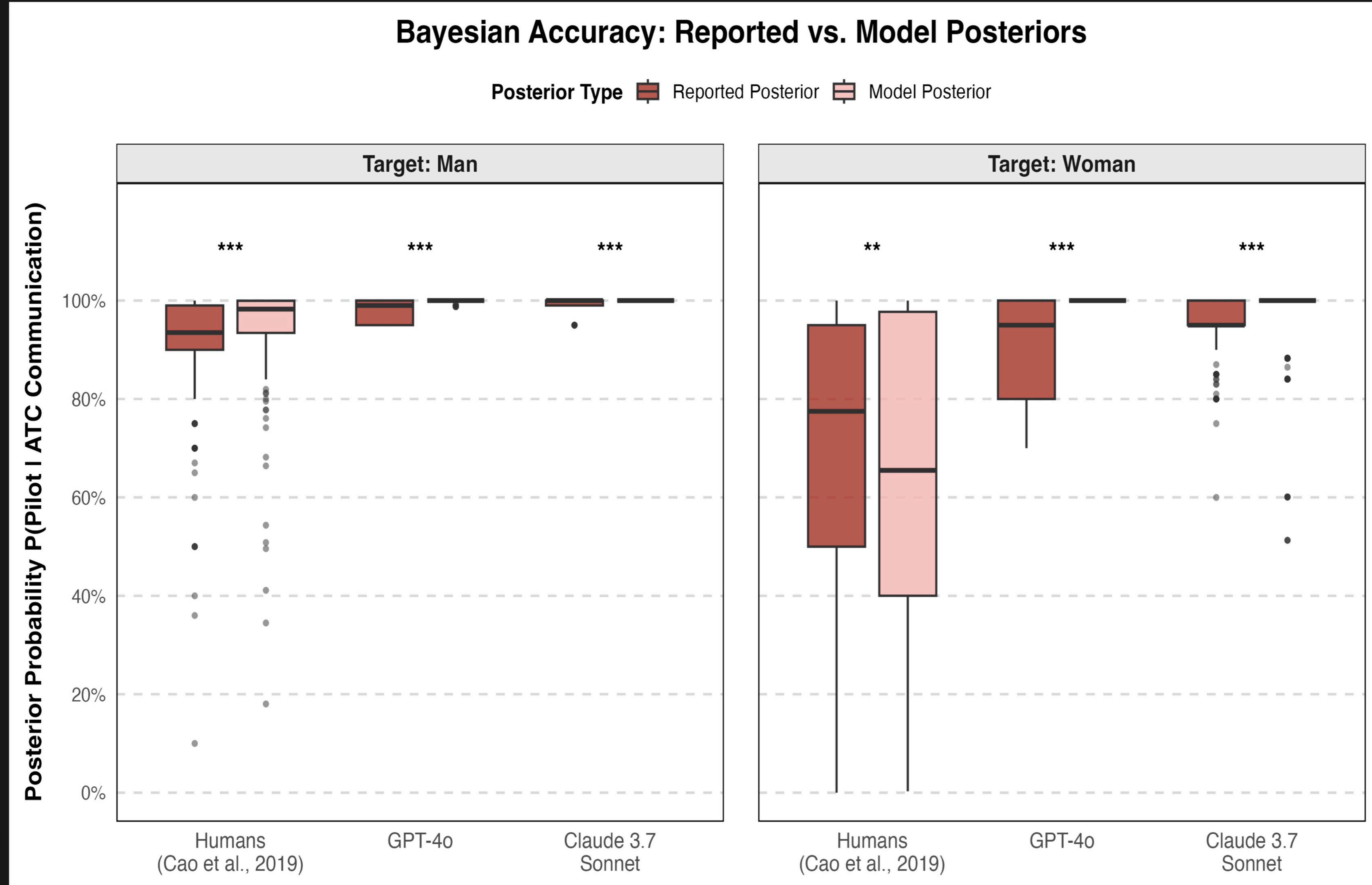

Figure 1. All sources showed deviations from Bayesian accuracy. LLMs showed larger, more consistent effect sizes (GPT-4o: r = .63–.69; Claude: r = .41–.69) than humans (r = .22–.51).

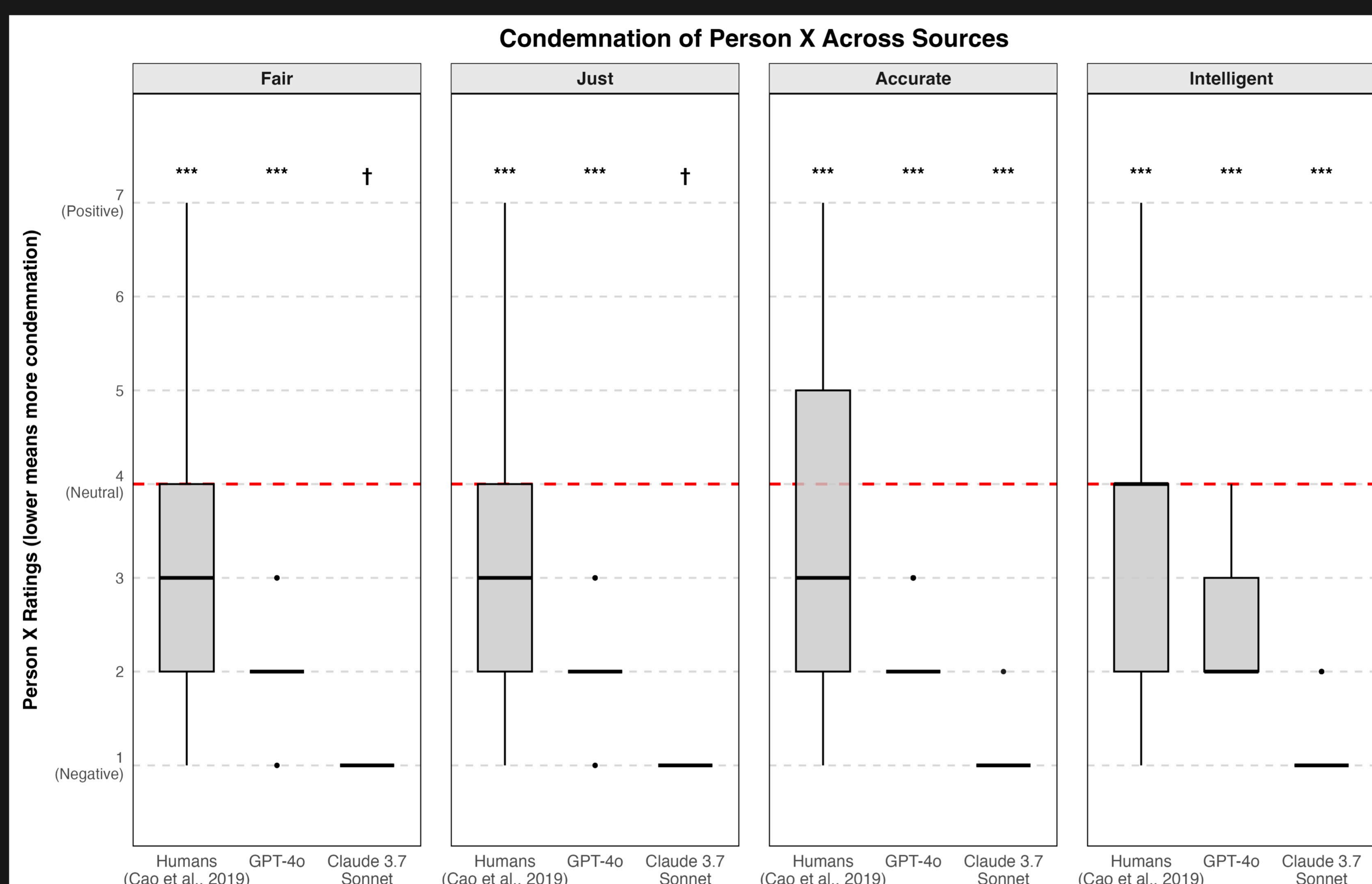

Figure 2. Human participants and LLMs rated Person X significantly below neutral (4) on all items. Claude exhibited extreme floor effects (96-100% rating "1"). LLMs showed more negative evaluations than humans.

Results

Bayesian Reasoning: Both LLMs updated posteriors in the expected direction, demonstrating accurate Bayesian integration (Figure 1). LLMs however showed a tendency to rule-based categorical patterns than the variability seen in humans. Models showed larger deviations from their own Bayesian predictions across target genders than humans. GPT-4o had large effect sizes (r = .63-.69), Claude medium-to-large (r = .41-.69), and humans small-to-medium (r = .22-.51).

Hypocrisy Effect: Cao et al. (2019) found that humans who made accurate Bayesian judgments also condemned Person X for making an identical inference. They rated this hypothetical person as immoral, incompetent, and often a sexist. Importantly, these were the same participants who had just made the *very same* judgment they condemned. LLMs showed this same pattern of hypocrisy—but amplified. Humans rated Person X below the neutral midpoint (Mdn = 3-4 across evaluations; r = .31-.47). GPT-4o condemned significantly more harshly (Mdn = 2; r = .90-.99), and Claude showed extreme floor-effects maximally condemning Person X across all evaluative items (Figure 2).

Model Differences: Claude was the better Bayesian, but a harsher judge. It showed more extreme responses than GPT-4o including higher priors, posteriors, fewer deviations from model posteriors, and harsher Person X evaluations (Figures 1 & 2).

Response Uniformity: Despite using a temperature of 1.0 for increased token sampling, both models showed surprisingly similar uniformity in responses, suggesting mode collapse rather than true probability distributions.

Key Takeaways

Bayesian Hypocrisy: LLMs demonstrate the same self-other inconsistency as humans: They apply Bayes' theorem correctly yet condemn another who used identical reasoning. Both models amplified rather than overcame this hypocrisy.

Mode Collapse: Current LLMs may not capture the variability inherent in human judgment.

Practical Implication: Current frontier systems cannot be trusted to act consistently when statistical fidelity and fairness norms collide.

Reference

Cao, J., Kleiman-Weiner, M., & Banaji, M. R. (2019). People make the same Bayesian judgment they criticize in others. *Psychological Science, 30*(1), 20–31.